

# Character Networks and Book Genre Classification

A.J. Holanda<sup>\*a</sup>, M. Matias<sup>b</sup>, S.M.S.P. Ferreira<sup>c</sup>, G.M.L. Benevides<sup>d</sup>, O. Kinouchi<sup>b</sup>

<sup>a</sup>*Departamento de Computação e Matemática - FFCLRP*

<sup>b</sup>*Departamento de Física - FFCLRP*

<sup>c</sup>*Departamento de Educação, Informação e Comunicação - FFCLRP*

<sup>d</sup>*Prefeitura do Campus USP de Ribeirão Preto*

*Universidade de São Paulo, Av. Bandeirantes 3900, CEP 14040-901, Ribeirão Preto, SP, Brazil*

---

## Abstract

We compare the social character networks of biographical, legendary and fictional texts, in search of statistical marks of historical information. We examine the frequency of character appearance and find a Zipf Law that does not depend on the literary genera and historical content. We also examine global and local complex networks indexes, in particular, correlation plots between the recently introduced Lobby (or Hirsh  $H(1)$ ) index and Degree, Betweenness and Closeness centralities. We also found no relevant differences in the books for these network indexes. We discovered, however, that a very simple index based in the *Hapax Legomena* phenomenon (names cited a single time along the text) that seems to have the potential of separating pure fiction from legendary and biographical texts.

**Keywords:** Social networks, Character networks, Lobby index, Hirsch index

---

<sup>\*</sup>Corresponding author.

E-mails: aholanda@usp.br (A. J. Holanda), okinouchi@gmail.com (O. Kinouchi)

## 1. Introduction

Social networks extracted from literary texts have been studied from some years to now. Most of the analyses characterized the networks of pure fictional texts with different indexes [1, 2, 3, 4, 5, 6]. Some of them intended to examine to what extent fictional social networks are similar or not to real (e.g. Facebook) networks [7, 8] and others proposed or test automatic social network extraction algorithms [9, 10].

Our aim was to perform a somewhat different approach. We compare social networks extracted from texts with pure fictional, legendary and biographical nature. From this corpus, we intend to find indexes that are capable to differentiate or classify pure fiction, legendary accounts with some historical traces and historical biography. Our main question is: legendary accounts are more like pure fiction or more like modern biographies?

In particular, we study a recent node centrality index, the Lobby index [11, 12], also called Hirsh  $H(1)$  index [13, 14]. We present correlation plots for the Lobby index versus Degree, Betweenness and Closeness centralities to check if it contains independent information that could be used to accomplish our task. We notice that even negative results are very useful, because they refute, in a Popperian way, the conjectures that network indexes can separate these genres.

## 2. Materials and Methods

We process the character networks from the following books: Bernard Cornwell's *The Winter King: a novel of Arthur* [15] (from here, *Arthur*), Charles Dickens *David Copperfield* [16] (*David*), J. R. R. Tolkien's *The*

*Hobbit* [17] (*Hobbit*), Mark Twain’s *Huckleberry Finn* [16] (*Huck*), *Luke Gospel* [18] (*Luke*), *Acts of the Apostles* [18] (*Acts*), Iamblicus’s *Life of Pythagoras* [19] (*Pythagoras*), James Gleick’s *Isaac Newton* [20] (*Newton*), and Humphrey Carpenter’s *Tolkien: a Biography* [21] (*Tolkien*).

We use an operational definition of fictional, legendary and biographical works. By pure fiction we denote an account that is recognized as such by the author of the book (*Arthur*, *David*, *Hobbit*, *Huck*). Legendary accounts are those that, in the view of modern scholars, contain fictional narratives mixed with possible biographical traces (*Lucas*, *Acts*, *Pythagoras*). Finally, biographical works are those recognized as such by modern standards (*Newton*, *Tolkien*).

We created the networks from the books with  $N$  characters represented by nodes and the characters encounters represented by links in an adjacency matrix  $A_{ij} \in [0, 1]$ . We gathered all network data manually, with the exception of *David Copperfield* and *Huckleberry Finn* that were obtained from the Stanford GraphBase project [16].

We calculated the following measures using NetworkX [22] Python library: Density  $D$ , average clustering coefficient  $\langle CC \rangle$ , node Degree  $K_i$ , node Betweenness  $B_i$  and node Closeness  $C_i$ . We also wrote Python scripts to evaluate the Lobby (or Hirsh  $H(1)$ ) index for node centrality [11, 12, 13, 14]. Additional information about the data and source code can be found at GitHub page for the project called CharNet<sup>1</sup>.

The density  $D$  of a graph is the ratio of the number of links and the

---

<sup>1</sup><https://ajholanda.github.io/charnet/>

possible number of links  $N(N-1)/2$ . The number of the node's neighbors is  $K_i$  and the network global Degree is  $K = 1/N \sum_i^N K_i$ . The clustering coefficient  $C_C$  is calculated as follows: coefficient:

$$\langle CC \rangle = \frac{1}{N} \sum_{i=1}^N \frac{2L_i}{K_i(K_i-1)}, \quad (1)$$

where  $L_i$  is the number of links between the  $K_i$  neighbors of node  $i$ .

The individual Degree  $K_i$  of a node can be used as a measure of its centrality and it can be normalized as  $K_i^N = K_i/N$ . Other centrality measures are Betweenness and Closeness. The Betweenness centrality  $B_i^N$  is defined as the number of shortest paths that pass through a node  $i$ , normalized by the number of pair of nodes not including  $i$ , that is  $(N-1)(N-2)/2$ . The Closeness centrality  $C_i$  is defined as the sum of shortest distances between a node  $i$  and all other reachable nodes, normalized to a maximum value  $C_i^N = 1$ . A character Lobby index is the maximum number  $L_i$  such that there exists at least  $L_i$  neighbors with degree larger than or equal to  $L_i$ , normalized as  $L_i^N = L_i/N$ .

Finally, we can study the frequency  $f_i$  that a given character name appear in the text. Notice that, due to operational reasons, we counted only explicit references to the name, not pronouns or indirect references to the character.

### 3. Results

*Character frequency.* We ranked the frequencies  $f_i$  in descending order, so that each character now has a rank  $R$  and a frequency  $F(R)$ . The plot  $F(R)$  is presented in Fig. 1. We normalize so that  $F(1) = 1$  and the horizontal axis is  $R/R_{\max}$ .

*Global indexes.* In Table 1 we show the global indexes average degree  $K$ , average clustering coefficient  $CC$  and density  $D$ .

<b>Book</b>	<b>N</b>	<b>Links</b>	<b>K</b>	<b>CC</b>	<b>D</b>
<i>Acts</i>	75	160	$4.27 \pm 5.19$	0.316	0.058
<i>Arthur</i>	108	155	$2.87 \pm 4.21$	0.005	0.027
<i>David</i>	87	406	$9.33 \pm 10.56$	0.351	0.109
<i>Hobbit</i>	65	161	$4.95 \pm 5.21$	0.000	0.077
<i>Huck</i>	74	301	$8.14 \pm 7.39$	0.488	0.111
<i>Luke</i>	76	203	$5.34 \pm 8.15$	0.340	0.071
<i>Newton</i>	46	44	$1.91 \pm 2.16$	0.000	0.043
<i>Pythagoras</i>	41	31	$1.51 \pm 2.20$	0.027	0.038
<i>Tolkien</i>	127	282	$4.44 \pm 9.14$	0.126	0.035

Table 1: Global network indexes for the books

*Node centrality indexes.* Remember that index  $i = 1, \dots, N$  enumerates individuals in a network with  $N$  characters. We examine correlation plots between classical centrality indexes (Degree, Betweenness and Closeness) and the recently introduced Lobby index. These are local centrality measures.

We plot in Fig. 2 the normalized Lobby index  $L_i^N$  versus the normalized degree  $K_i^N$  for all characters (some graphs, as *Pythagoras*, show few points because they have the same  $(L_i^N, K_i^N)$ ). We can see that there is an initial linear correlation between the Degree and Lobby indexes followed by a saturation. This can be understood because it is much harder for the Lobby index to grow after some point. For example, it is possible for the central character to have degree  $D_i^N = (N - 1)/N \approx 1$  (he/she knows all the other

personages) but for having  $L_i^N \approx 1$  we would need an all-to-all (complete) graph where not only the central character knows all the people, but also any of his/her neighbors also knows all the people.

Comparing all the nine plots, it seems that they are mostly equal and correlation between Lobby and Degree cannot separate the book genres. See, for example, the plots for *David*, *Huck*, *Luke* and *Tolkien*, which are almost indistinguishable.

The Pearson correlation is weak between Lobby and Betweenness (Fig. 3). We also notice that the correlation is greater for the biographies than for most of the fictional and legendary texts. However, the fictional book *Arthur* has a larger correlation than *Tolkien*, so we have a counterexample for that trend.

We observe an interesting phenomenon in the Lobby vs Closeness plot (Fig. 4). It shows clusters in the data, a feature already found in a study of biological networks [12]. It seems that Lobby can detect clusters or communities that the other indexes cannot detect. However, these clusters appear in *Huck*, *David*, *Luke* and *Tolkien*. So, anew, these correlation plots cannot separate the book genres.

We notice that the *Pythagoras* plot is very poor because several characters have the same Closeness. The *Pythagoras* character network has a low number of links when compared with other networks used in the study. However we maintained *Pythagoras* in our sample because it is a prime example of legendary account. It also gives us an idea about the behavior of books with small number of characters.

*Hapax Legomena*. Finally, we found a very simple measure that has the potential of distinguishing the books. From literary criticism we have that

words that appear a single time in a text are called *Hapax Legomena*. Here we consider only *Hapax Legomena* (HL) for character names, that is, names with frequency  $f_i = 1$ . They are presented in Table 2, with the books ranked from the largest to the lowest *Hapax Legomena* ratio  $HL^N = HL/N$  (number  $HL$  of names with  $f_i = 1$  divided by total number of characters  $N$ ). We also report *Dis Legomena* (DL) names, that is, names with  $f_i = 2$ .

<b>Book</b>	<b>HL<sup>N</sup> = H/N</b>	<b>DL<sup>N</sup> = DL/N</b>
<i>Newton</i>	41/46 = 0.891	14/46 = 0.304
<i>Acts</i>	51/75 = 0.680	13/75 = 0.173
<i>Luke</i>	51/76 = 0.671	15/76 = 0.197
<i>Tolkien</i>	65/127 = 0.512	28/127 = 0.220
<i>Pythagoras</i>	21/41 = 0.512	08/41 = 0.195
<i>Arthur</i>	52/108 = 0.481	20/108 = 0.185
<i>Huck</i>	32/74 = 0.432	19/74 = 0.257
<i>David</i>	26/87 = 0.299	09/87 = 0.103
<i>Hobbit</i>	18/65 = 0.277	10/65 = 0.154

Table 2: Number of character names that are *Hapax legomena*  $H$  divided by total number  $N$  of characters. The books have been listed in descending order for  $H^N$ .

#### 4. Discussion

The task to distinguish real social and purely fictional networks is a hard one [7, 8]. The issue complicates when we study legendary texts, which we define as text that cannot be trusted as historical biographies but could have some historical traces due to oral traditions. we have no certainty that

the social network described is fictional or some information refers to true historical social relations. This is the case of the narratives about Pythagoras, Jesus of Nazareth and the first apostles.

Indeed, the normalized frequency  $F(R)$  of name citations follows a Zipf law (for character appearance, not words!), it is universal and does not depend on the literary genera examined (see Fig. 1). Of course this statement needs to be confirmed with a larger corpora, but anyway it suggests that  $F(R)$  is not a good measure to distinguish historical from fictional accounts.

In the case of the global measures as average degree, density and average clustering coefficient (Table 1), we see no trend that separates the genres. This result suggests that these global measures are not good metrics to classify the texts, they depend on the size and structure of the books, a conclusion already advanced in [7, 8].

Global measures may be used to analyze the elements of literary narrative like social importance, psychological depth, sociological breadth, weight of social ties, character interaction, egocentric focus on some character, nature of relationships, among others [23].

Recently, Ronqui and Travieso [24] proposed that the analysis of correlations between centrality indexes is interesting to characterize and distinguish between natural and artificial networks. We examined the correlation plots for the Lobby index versus Degree (Fig. 2), Betweenness (Fig. 3) and Closeness (Fig. 4). Such plots revealed that social networks, fictional and legendary or historical are very similar and they cannot be easily distinguished.

Although these are negative results, we thought that they are important ones. After all, with such small sample, we cannot aim to have corroboration



by induction (a large number of results suggesting some conclusion). Indeed, even with perhaps a sample of one thousand books, nothing prevents that the next one (or the next thousand ones) refutes the conclusions. On the other hand, negative results refute conjectures, as Popper so clearly showed. And, indeed, our small sample refutes a lot of *a priori* conjectures concerning the capacity of traditional network indexes to separate the genres.

However, another idea could be that there is no clear motivation for a writer of a pure fictional work to introduce a character in a single scene, and cite his/her name only one or two times. Such constraint is weaker for biographies, where characters appear due to historical events and not from the special creative work of the author. This is also valid for legendary accounts, that are more fragmented and follows ancient writing styles. So, our hypothesis is that the presence of *Hapax legomena* for characters would be more rare in pure fiction.

This hypothesis is confirmed by our data, there is a clear trend in Table 2 where pure fictional works have less *Hapax Legomena* per character, in the interval  $[0.277, 0.481]$ ; modern biographies are in the interval  $[0.512, 0.891]$ ; and ancient legendary accounts lie in the interval  $[0.512, 0.680]$ . This means that a value of 0.5 separates pure fiction from the other books. We think that this preliminary result, which distinguishes legendary and historical works from pure fiction, is important and makes intuitive sense. This trend, however, is not observed for the *Dis Legomena* data. Of course, this result must be confirmed by a larger study but, by now, the idea about *Hapax Legomena* seems to be promising and has not been refuted.

## 5. Conclusion and Perspectives

We examined three questions in the current research: first, is there some difference among pure fictional social networks (centered in a main character), legendary social networks and networks extracted from a historical biography? Second, are there complex network indexes with potential to separate these genres? Third, what is the behavior of the recently introduced Lobby index in this respect?

This first study is important by posing the problem and exploring its possible solutions. Even with a small sample, our findings seems to refute some ideas like to use the name frequency  $F(r)$  and global measures as average degree, density and average clustering coefficient as discriminators.

By examining local node centrality indexes like Degree, Closeness, Betweenness and Lobby, what we obtain is that to separate the genres by using only the social networks is a hard and non trivial task. Although negative, these results are important as guide for future research.

To overcome the limitations of this paper, we foresee only a (non trivial) methodological advance: to have a good algorithm that extracts automatically social networks from raw texts. Since this methodology is yet under development [9, 10], our study can be thought as both preliminary and as a benchmark for further studies.

Although our literary corpus is very small, the current work intends to present a methodology for the study of an old question on how to extract historical information from legendary accounts to examine character networks. Our work also suggests the use of the fraction of *Hapax Legomena* for characters to separate the fictional texts from the legendary and biographical ones.

This result seems to be very simple but important and it may reflect the fact that legendary accounts could have some traces of true historical social networks.

## **Acknowledgments**

This paper results from research activity on the FAPESP Center for Neuromathematics (FAPESP grant 2013/07699-0). OK acknowledges support from Núcleo de Apoio à Pesquisa CNAIPS-USP. MM received support from PUB-USP.

## **Author contributions statement**

GMLB, SMSPF, MM and AJH extracted the books character networks and character frequency data. AJH organized the public database, performed the complex network analyses and analyzed the data. OK proposed the original problem and analyzed the data. AJH and OK wrote the paper. All authors reviewed the manuscript.

## **Competing financial interests**

The authors declare no competing financial interests.

## **References**

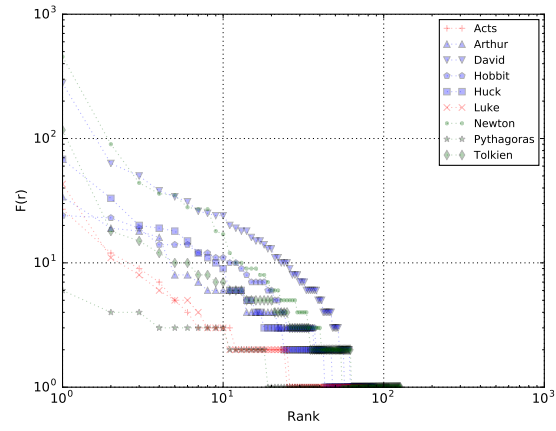
## **References**

- [1] Y.-M. Choi, H.-J. Kim, A directed network of greek and roman mythology, *Physica A* 382 (2) (2007) 665–671.

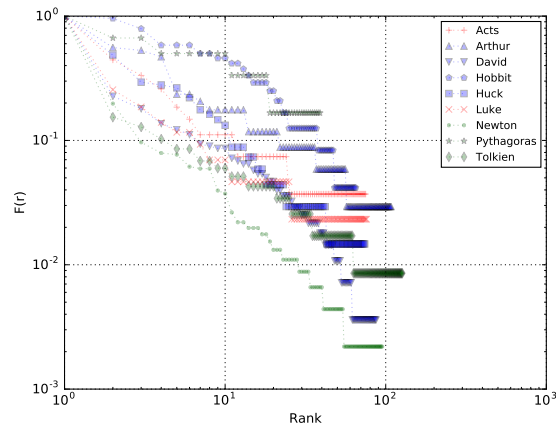
- [2] P. Mac Carron, R. Kenna, Universal properties of mythological networks, *EPL (Europhys Lett.)* 99 (2) (2012) 28002.
- [3] A. Agarwal, A. Corvalan, J. Jensen, O. Rambow, Social network analysis of alice in wonderland, in: *Workshop on Computational Linguistics for Literature*, 2012, pp. 88–96.
- [4] P. Mac Carron, R. Kenna, A quantitative approach to comparative mythology, *Cosmos* 14 (2013) 103–117.
- [5] R. Kenna, P. Mac Carron, Maths meets myths: Network investigations of ancient narratives, in: *Journal of Physics: Conference Series*, Vol. 681, IOP Publishing, 2016, p. 012002.
- [6] M. A. Ribeiro, R. A. Vosgerau, M. L. P. Andruchiw, S. E. d. S. Pinto, The complex social network of the lord of rings, *Rev. Bras. Ens. Fís.* 38 (1).
- [7] R. Alberich, J. Miro-Julia, F. Rosselló, Marvel universe looks almost like a real social network, *arXiv preprint cond-mat/0202174*.
- [8] P. M. Gleiser, How to become a superhero, *J. Stat. Mech. Theor. Exp* 2007 (09) (2007) P09020.
- [9] D. K. Elson, N. Dames, K. R. McKeown, Extracting social networks from literary fiction, in: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2010, pp. 138–147.

- [10] S. Grayson, K. Wade, G. Meaney, D. Greene, The sense and sensibility of different sliding windows in constructing co-occurrence networks from literature, in: Computational History and Data-Driven Humanities: Second IFIP WG 12.7 International Workshop, CHDDH 2016, Dublin, Ireland, May 25, 2016, Revised Selected Papers 2, Springer, 2016, pp. 65–77.
- [11] A. Korn, A. Schubert, A. Telcs, Lobby index in networks, *Physica A* 388 (11) (2009) 2221–2226.
- [12] M. G. Campiteli, A. J. Holanda, L. D. Soares, P. R. Soles, O. Kinouchi, Lobby index as a network centrality measure, *Physica A* 392 (21) (2013) 5511–5515.
- [13] L. Lü, T. Zhou, Q.-M. Zhang, H. E. Stanley, The h-index of a network node and its relation to degree and coreness, *Nat. Commun* 7 (2016) 10168.
- [14] R. Pastor-Satorras, C. Castellano, Topological structure and the h index in complex networks, *Phys. Rev. E* 95 (2) (2017) 022301.
- [15] B. Cornwell, *The winter king: a novel of Arthur*, Penguin, London, 2007.
- [16] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*, ACM Press and Addison-Wesley, 1994.
- [17] J. Tolkien, *The Hobbit*, Houghton Mifflin Harcourt, Boston, 2012.
- [18] American Holy Bible, *The Revised American Standard Version Bible Society*, 2011.

- [19] T. Taylor, Iamblichus' Life of Pythagoras, Inner Traditions/Bear & Co, Rochester, 1986.
- [20] J. Gleick, Isaac Newton, Vintage Books, New York, 2004.
- [21] H. Carpenter, J.R.R. Tolkien: a biography, Houghton Mifflin Harcourt, Boston, 2014.
- [22] A. A. Hagberg, D. A. Schult, P. J. Swart, Exploring network structure, dynamics, and function using NetworkX, in: Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA USA, 2008, pp. 11–15.
- [23] G. A. Sack, Character Networks for Narrative Generation: Structural Balance Theory and the Emergence of Proto-Narratives, in: M. A. Finlayson, B. Fisseni, B. Löwe, J. C. Meister (Eds.), 2013 Workshop on Computational Models of Narrative, Vol. 32 of OpenAccess Series in Informatics (OASISs), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013, pp. 183–197.
- [24] J. R. F. Ronqui, G. Travieso, Analyzing complex networks through correlations in centrality measurements, J. Stat. Mech. Theor. Exp 2015 (5) (2015) P05030.



(a) Non-normalized



(b) Normalized

Figure 1: Zipf curves for character appearance frequency.

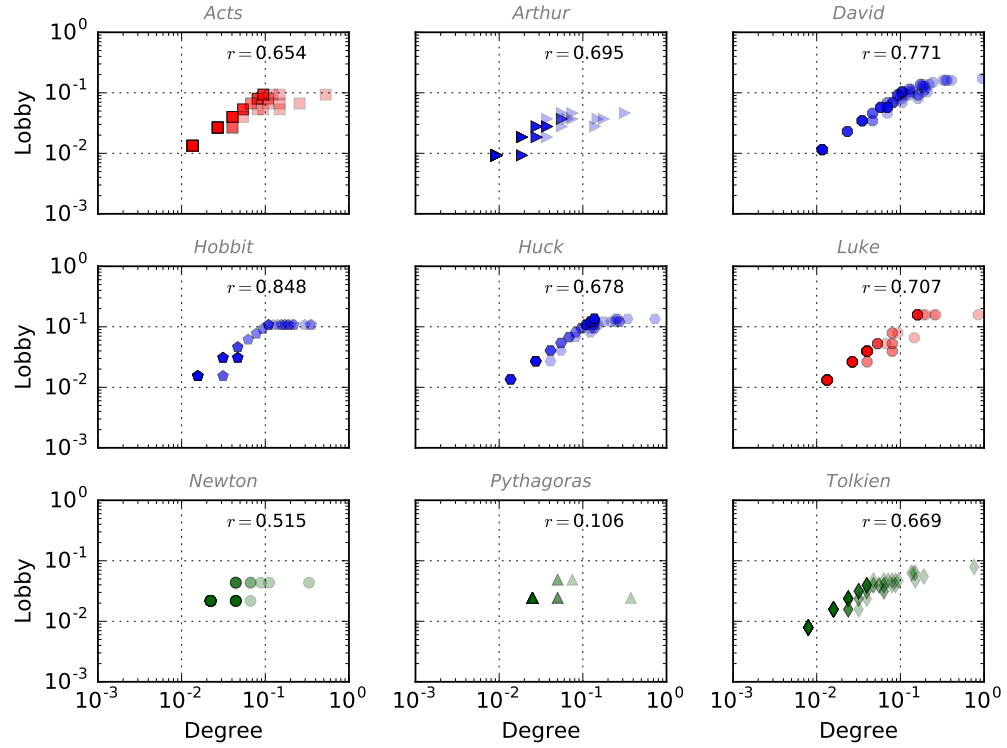


Figure 2: Correlation plots for Lobby versus Degree centrality with Pearson correlation  $r$  at the top.



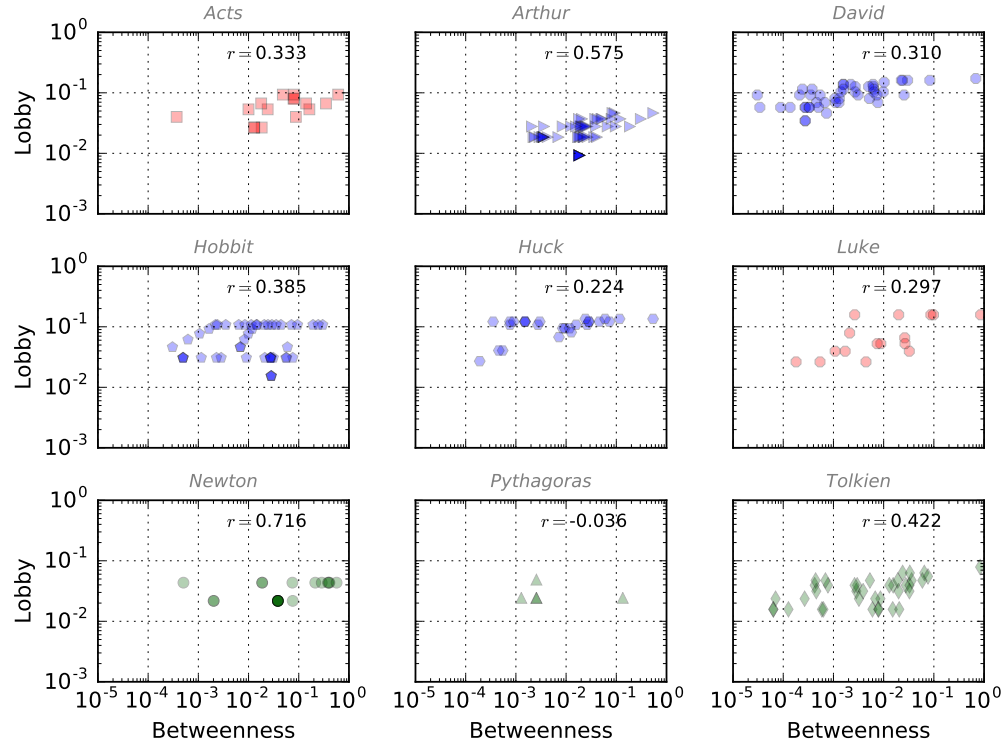


Figure 3: Correlation plots for Lobby versus Betweenness centrality with Pearson correlation  $r$  at the top.

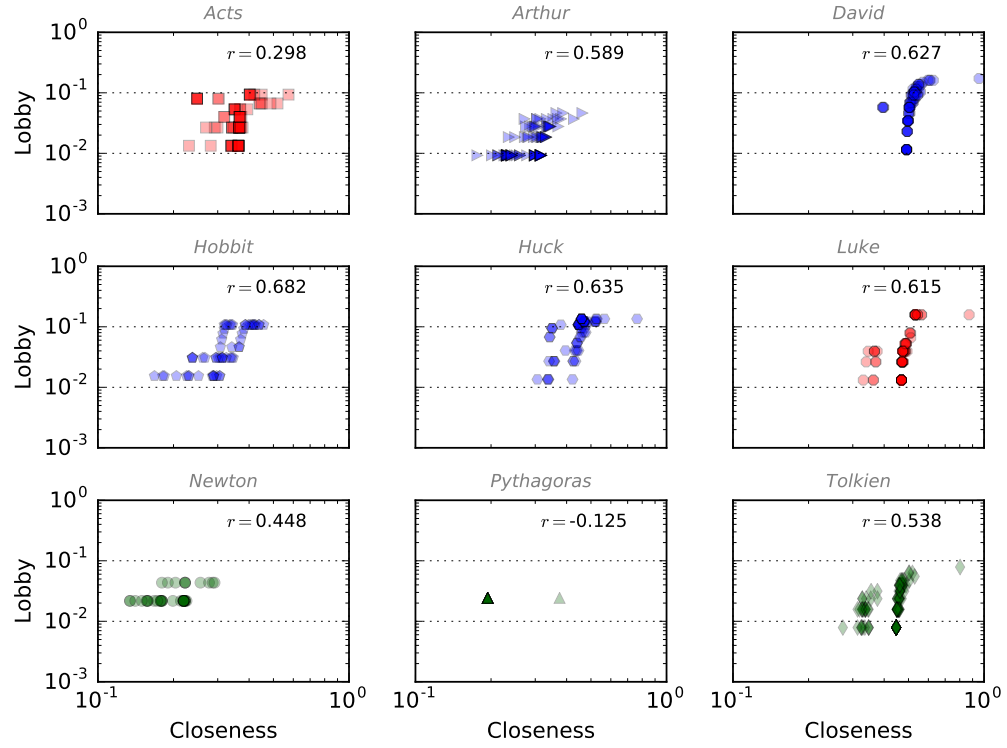


Figure 4: Correlation plots for Lobby versus Closeness centrality with Pearson correlation  $r$  at the top.